Validating Patient Names in an Integrated Clinical Information System

Robert V. Sideli, M.D.
Carol Friedman, Ph.D. *

Columbia-Presbyterian Medical Center, New York
* Queens College of the City University of New York

Methods for validating patient names during the upload of clinical records are described. Exact string matching, Soundex method and a pattern matching algorithm (LCS method) are described and compared to a manual analysis of 10000 patient name pairs. In addition, the types of spelling and typographical errors that occur in patient names in the pathology database at CPMC are described. The data analysis shows that the LCS method performs better than the other techniques when compared to manual analysis.

Introduction

Columbia-Presbyterian Medical Center (CPMC) is building an integrated Clinical Information System (CIS) with a centralized database. Clinical patient data is added to the CIS database by processes which upload data from departmental systems [1]. An additional component of the CIS is a patient registry, which contains demographic patient data and associates unique patient identification numbers (PID) with each patient. The patient is identified throughout the medical center with this unique PID. While many of the medical center's departmental systems are connected to an institution wide network [2], there is little if any validation of the patient information in the departmental systems. Most of the departmental systems maintain their own patient and result databases. In the best of all worlds, the departmental systems could continue to maintain their own databases, but there would be an ongoing process that would synchronize the patient identity across these systems [3]. While we could technically accomplish this synchronization between the CIS and a few departmental systems, it would be a large initial impediment to collecting clinical results. Therefore, we sought an alternate solution.

Initial studies showed that discrepancies in patient identification are very common (20-30%) and thus, validating patient names poses a considerable problem to the CIS system developers. One approach would be to allow all the patients with valid PIDs to have their clinical records uploaded into the result database. At the time the clinical user reviews the patient's test result, the patient name in the upload record could be displayed with the test results. The clinical user would then exercise his or her own judgement to decide whether to accept the results as that of their patient. This was the approach that was used for many years in the Laboratory Information System of Presbyterian Hospital and was continued when the reporting of clinical laboratory results became available in the CIS. However, as new report types (e.g. radiology reports, pathology reports, discharge summaries, etc.) became available through the CIS it was felt that patient care could potentially be compromised if the clinical user incorrectly attributed a test result to the wrong patient. In fact, we learned of several incidents where physicians asked their patient's about recent tests that they had, only to be informed by the patient that they never had any such tests. It was felt that we needed a reliable means of validating patient names during the upload of clinical results. A method was needed that would alleviate the clinical user of the burdensome and error prone task of constantly having to validate the patient identification of test results. This method needed to perform as well as the perceptive clinical user. We report here our experiences with several name validation methodologies.

Methods

The test set of data consists of the patient name and PID from 10,000 unique patient entries in the Department of Pathology data management system. The PIDs of the patient records were used to extract the corresponding patient names from the CIS patient registry. The resulting name pairs were then subjected to several comparison methods.

The data analysis consists of construction of 2x2 tables where the cells are true positive, false positive, false negative and true negative. The results of the below described comparison techniques are each compared to the results of the manual analysis ("gold standard"). The sensitivity, specificity, false positive rate and false negative rate are then calculated. Since the objective of

This work was supported in part from a grant from the National Library of Medicine LM04419 (IAIMS), and LM00002-01

the comparison techniques is to find true patient name mismatches, a mismatch by one of the methods is considered a positive result. A true result occurs when both the manual method and test being studied yield similar results. To minimize confusion, all references to positive and negative will be recast as mismatches and acceptances. For example, a false positive will be called a false mismatch and a false negative will called a false rejection.

The date of birth and sex are usually recorded in both the pathology and CIS patient registry. Unfortunately, even when present the quality of these data is poor and therefore, the comparison methods listed below are limited to the patient name.

Exact String Match and Manual Analysis

An exact string match was performed on the name pairs (upload record and CIS patient registry) and the output was subjected to a manual analysis to determine the true positive mismatch rate. The person (RVS) manually comparing the name pairs was very familiar with the hospital environment and the ethnicity and history of the patient population, and therefore used this background knowledge when matching names. A manual analysis of 560 randomly chosen pathology name mismatches was also performed. This was done in order to categorize the types and frequency of errors in the patient names.

Soundex Method

Algorithms commonly used for correcting or tolerating name misspellings are typically based on the Russell Soundex Code [4-7] which accepts names that sound alike. It is one of the best known methods which performs phonetic reductions on names. The underlying principle is that names that sound alike should reduce to the same code.

The Soundex code for a name as given by Knuth [7], consists of the initial letter of the surname plus three digits derived from the remaining letters of the surname as follows:

- All vowels, and the letters H, W, and Y are dropped
- The following letters compute to the following digits:
  1 - B,F,P,V
  2 - C,G,J,K,Q,S,X,Y,Z
  3 - D,T
  4 - L
  5 - M,N
  6 - R
- All consecutive repeating digits are ignored
- If there are less that three digits add trailing zeros so that the codes consists of a letter and exactly three digits.
- For example, Sideli = S340

A maximum of 26*7*7*7 (8918) different codes can be obtained. This means that there are usually many names corresponding to one particular code. Using this method, the Soundex code for each member of the name pair was calculated and the values compared. The pairs with different numbers were considered name mismatches by the Soundex method.

Longest Common Substring Method

This technique has been adapted from an algorithm developed by Baskin and Selfridge [8]. We call it the Longest Common Substring (LCS) method because it is based on the notion of a likeness measure between two strings. The likeness is obtained using a procedure which iteratively finds and removes the longest common substring between the two strings. The common substring must be longer than a minimum length threshold limit, which was set to three for this study The likeness measure is based on the total length of the common portions of the name pairs compared to the length of the actual names. This measure can be calculated by dividing the total length of the common portions by the length of the smallest of the two strings.

The technique can be demonstrated by the following examples. In the example, we use the name pair *Bobby Huntington* and *Robbie Huntinton*

1. An LCS *Huntin* of length 6 is found and removed from both names, leaving a pair of strings *Bobby gton* and *Robbie ton.*
2. An LCS *obb* of length 3 is found and removed leaving the pair *y gton* and *R ie ton.*
3. An LCS *ton* of length 3 is found and removed leaving the *y* and *R ie.*
4. There are no more common substrings, and the iteration procedure ends.

The common portion of the two substrings has a total length of 12 (6+3+3). Dividing the length of the common portion by the length of the smallest of the two strings gives the likeness measure, which in this case is 80%.

Results

When the PIDs of 10,000 unique patient entries in the pathology database were searched for in the CIS database 3.4% of the pathology PIDs were not found in the CIS patient registry. The data analysis (table 1) is based on the number of patient names actually found (9663) in the CIS patient registry.

It should be remembered that the following description considers that a mismatch by one of the methods is a positive result.

Requiring an exact string match between the upload name and patient registry name results in a mismatch rate of 22.4%. When the exact string match rejects are manually reviewed, the mismatch rate is 1.6%.

The Soundex method of comparing names resulted in a mismatch rate of 6.2%. There were no false acceptances. This is a high mismatch rate when compared to the manual analysis, but there is about a 70% reduction in the mismatch rate when compared to the exact string match test. There were 440 false mismatches.

Table 1 shows the results of applying the LCS method, with the likeness threshold varying from 0.30 to 0.60. The mismatch rate varies from 1.5% to 3.0% respectively. The absolute number of false mismatches range from 10 to 136 and the number of false acceptances range from 19 to 1.

Manual inspection of the 560 randomly chosen name pair mismatches from the pathology database reveals that the most frequent types of errors in patient names are categorized as follows:

1. Insertion/deletion of additional names, initials and titles (36.4%)
   Smith, Mary; Smith, Mary Ann
   Smith, John; Smith, John Jr.
2. Several letters of the name are different due to nicknames and slight spelling variations (13.9%)
   Nicholas; Nick; Nickie; Nicky
3. One letter is different (13.7%)
   Nicholas; Nickolas
4. One letter added or deleted (12.9%)
   Gomnez; Gomez
5. Differences due to punctuation marks and number of blanks (11.8%)
   O'Connor; O Connor; OConnor
6. Different last name for female patient (7.8%)
   Gomez, Ann; Vega, Ann
   Note: Approximately one half of the total pathology records come from cytopathology (PAP smears) and thus, there is a higher proportion of female patients than expected.
7. Parts of the name are permuted (1.4%)
   Gomez, Ann; Ann Gomez
8. Different first name (1.4%)
   Smith, Helen; Smith, Ellen
9. Permutation of 1 letter (0.8%)
   Robrets, Bill; Roberts, Bill

The frequency rates of errors in patient names are based on the 591 errors found. There are more errors than mismatches (560) because occasionally more than one error is associated with a mismatched pair.

Manual analysis also revealed an interesting procedure in the patient registry. Apparently when the Medical Records Department discovers that a patient has a duplicate PID, the name in the registry is changed to the character string "Use PID: xxxxxxx". This works fine for on-line users, however, our study revealed 28 such occurrences and they were rejected by all methods. This practice has some merits in that the name mismatches are routed back to the sending department and they are shown the two names. In this situation, they could simply change the PID and resend the report.

Discussion

This study is based on the environment of CPMC. Several factors are present at CPMC that may not be typical of other healthcare facilities, and therefore our

Table 1

	TP	FP	FN	TN	Sens.	Spec.	FNR	FPR
Exact	158	2003	0	7502	1.0000	.7893	.0000	.2107
Soundex	158	440	0	9065	1.0000	.9537	.0000	.0046
LCS-30	139	10	19	9465	.8797	.9989	.1203	.0011
LCS-35	146	10	12	9495	.9241	.9989	.0759	.0011
LCS-40	151	12	7	9493	.9557	.9987	.0443	.0013
LCS-45	154	27	4	9478	.9747	.9972	.0253	.0028
LCS-50	155	33	3	9472	.9810	.9965	.0190	.0035
LCS-55	157	91	1	9414	.9937	.9904	.0063	.0096
LCS-60	157	136	1	9369	.9937	.9857	.0063	.0143

TP - true positive, FP - false positive, FN - false negative, TN - true negative
Sens. - sensitivity, Spec. - Specificity, FNR - false negative rate, FPR - false positive rate

study may be valid only in similar environments. At CPMC, one factor that probably increases the frequency of errors in patient names is that the patient population consists of a broad range of ethnic diversities. It is generally more difficult for a native speaker to spell a name correctly which is of foreign origin. Another factor that probably effects the type and frequency of errors is that CPMC is a large-scale busy facility with a transient patient population and therefore, the names are generally unknown to the departmental personnel that handle record keeping.

The manual analysis performed by one observer served as the "gold-standard" test in our study. This observer (RVS) is one of the authors, however, the LCS algorithm was devised by the second author (CF). While this is not an optimal situation, the manual analysis was performed blind to the results of the Soundex and LCS methods. An additional study should be performed by an independent auditor to validate the current manual analysis. Even if all sources of bias could be removed from the manual analysis, using this method as a "gold-standard" is still not without problems. While the manual analysis replicates the procedure performed by the health care provider during results review, the only true "gold-standard" would be the real rate of patient name mismatch. It would be nearly impossible to contact each patient in some study set to validate their identity. However, one could study several clinical and demographic attributes that are stored in each system (e.g. date of birth, sex, clinical diagnosis, etc.). This study, while costly and time consuming should be able to furnish the prior probability that given two patient names, the odds that they represent the same individual. It should be noted that in this current study we have ignored the possibility, that two identical names with the same PID could actually represent two different individuals.

While the exact string match method has a sensitivity of 100%, its specificity (78.9%) is the lowest of all the methods we have tested. Since our objective is to include the greatest number of cases, it is obvious that a false mismatch rate of 21% is unacceptable.

The Soundex method performed similarly to the exact match in that there were no false acceptances, but it was better than the exact string method in that the false mismatch rate was lower. However, the false mismatch rate of 4.6% is still unacceptably high since our primary objective is to include the greatest number of cases. A significant portion of these false mismatches is due to females with different last names. Soundex methods are usually applied to only the patient's last name. However, the study of 560 pathology names shows that 7.8% of the name errors were due to different last names for female patients. This analysis also reveals a 1.4% rate of permutation of parts of complete name. These sources of errors could be compensated for by performing Soundex coding on both the last and first name and accepting either code. Another shortcoming of most Soundex methods is the fact that the first character of the last name must match exactly and therefore, any errors in this character guarantees failure.

Soundex coding was developed to perform sound-alike comparisons and is typically used when data entry is being done with no written version of the name. In our case, the departments receive requisitions which have the patient name and PID. Any errors that occur while transferring this information into the departmental system are mainly due to typographical errors and the Soundex coding schema does not correct for common typos. Therefore it is not surprising that this method has such a high false mismatch rate. The Soundex method has found its widest application in on-line name lookups. In this situation, a secondary index is maintained and the key is the Soundex code of a person's name. When a name is searched in the database, it is first converted to its Soundex code and that is used as a hash code to find all the names that result in the same Soundex code. This is a particularly efficient means of performing a name search that tolerates a marked degree of misspellings.

With the LCS-60 method there is further improvement in the specificity (97.4%), however, the false mismatch rate is still fairly high (1.4%). There was 1 false acceptances, which could be explained by the algorithm. It occurred when the name pair consisted of 2 male patients with the same last name and short first names.

The LCS methods around the 0.40 threshold come the closest to the manual analysis. In fact, with LCS-40 the specificity rises to 99.9%, and the number of false mismatches drops to 12 with a false mismatch rate of 0.13%. However, for the first time there is a significant increase in the number of false acceptances. With the LCS-40 method there are 7 false acceptances with a false acceptance rate of 4.4%. These are reports that the manual analysis considered mismatches, but that the LCS method tolerated. These false acceptances consisted of males with the same last name and short first names, females with different last names but with the same name and not the same date of birth (i.e. not a maiden name), patients with the same last name, and finally a short last name that is a fragment of the other member of the pair.

If the algorithm does not tolerate the name error the clinical record is not uploaded and the record is sent back to the department and division of origin. The true mismatches signify that violations of the integrity of the

patient databases were prevented. The department personnel receive an error report which shows the name mismatch and they manually correct the report and resend the report. False mismatches need to be minimized because the clinical data will not be available for review by the clinical users of the CIS and additionally, the false mismatches will be returned to the department of origin and this represents an extra burden on the departmental personnel.

Currently, the CIS does not display results in summary mode and therefore, the user is always shown at the time of data review, the patient name which appeared on the test requisition. However, when summary reporting is implemented, this information will move to a detail screen for the individual tests and the user will have to explicitly ask for it. Additionally, CPMC has implemented an automated decision support system [9] and as more data passes through this system, false acceptances will start to take on increased importance. The designers of the decision support system need to consider the possibility of false acceptance of patient records. The other possibility is that to guarantee a minimal false acceptance rate we will have to increase the likeness threshold, which unfortunately will most likely be accompanied by an increase in the false mismatch rate. However, as was stated earlier, until be have a "true gold-standard" for our institution we really don't know how often two different patients have the same name and PID.

Difficulties in linking patient records has great implications for the health care community in general. A recent report to Congress discusses the feasibility of linking research and administrative databases [10]. The reports states that "the greatest technical impediment to linking personal data files is the lack of a standard identifier". One proposal is to use a set of uniformly collected variables that uniquely identify a person, for example, name, date of birth, sex, ZIP Code of residence , street address. In fact, many government data collectors (e.g., Medicare claims, registries, etc.) link record using names, addresses, and/or hospital specific medical record numbers and other variables. Unfortunately, within a given hospital, the ancillary departments usually do not store or request from the patient the additional identifiers which could uniquely identify a patient.

Conclusion

Our study has shown that an algorithm based on a likeness measure, when compared to manual comparison, is a good method for tolerating errors in patient names. Because the LCS method removes the common substring from each name of the pair and repeats the matching process, it is effective in handling a variety of different type of errors, such as minor typos, small variations, and name permutations. The current implementation of the LCS-40 method rejected only 0.05% more name pairs than the manual analysis, but had a significant false acceptance rate of 4.4%. Since we are currently considering the clinical users as the final arbitrator of the correctness of the patient name pair, we are willing to accept these false acceptances while we explore modifications of the methodology that will minimize this rate.

References

[1] Sideli R, Johnson S, Weschler M, Clark A, Simpson R, Chen C. Adopting HL7 as a Standard for the Exchange of Clinical Text Reports. Proceedings Fourteenth Symposium on Computer Applications in Medical Care, edited by Miller RA, New York, New York, IEEE, pp. 226-229, 1990.

[2] Sengupta S. Heterogeneity in Health Care Computing Environments. Proceedings Thirteenth Symposium on Computer Applications in Medical Care, edited by Kingsland LC, New York, New York, IEEE, pp. 355-359, 1989.

[3] Hammond WE, Straube MJ, Stead WW. The Synchronization of Distributed Databases. Proceedings Fourteenth Symposium on Computer Applications in Medical Care, edited by Miller RA, New York, New York, IEEE, pp. 345-349, 1990.

[4] Wiederhold GW. Databanken, volume 2. R. Oldenbourg Verlag, Munchen, Wien, 1980.

[5] Russell RC, April 2 1918. U.S. Patent 1,261,167.

[6] Russell RC, November 14 1922. U.S. Patent 1,435,663

[7] Knuth DE. The Art of Computer Programming, volume 3. Addison Wesley, Reading, Mass, 1973.

[8] Alberga CN, String similarity and misspellings. CACM, 10:302-313, 1967

[9] Hripcsak G, Clayton PD, Cimino JJ, Johnson J, Friedman C. Medical Decision Support at Columbia-Presbyterian Medical Center. YMHA Working Conference on Software Engineering in Medical Informatics, Amsterdam, The Netherlands, 8-10 October 1990.

[10] Office of Science and Data Development, AHCPR. The Feasibility of Linking Research-Related Data Bases to Federal and Non-Federal Administrative Data Bases: Report to Congress. Rockville (MD): DHHS, PHS, Agency for Health Care Policy and Research (AHCPR) Center for Research Dissemination and Liaison; 1991 May. Report No.: AHCPR 91-25,